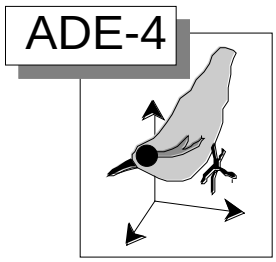


PCA

PCA : After row % transformation PCA

Méthode d'analyse multivariée à un tableau.

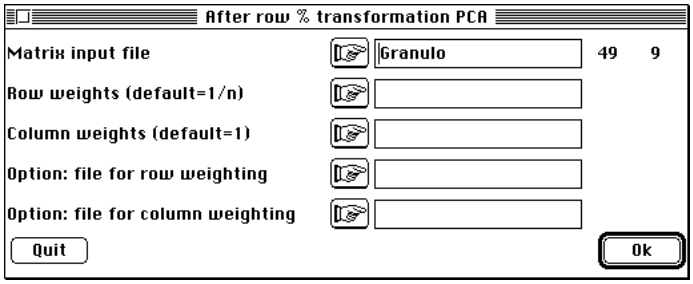
L'analyse en Composantes Principales (ACP) sur les profils lignes est l'analyse d'un schéma de dualité avec le jeu de paramètres :

1 — Transformation initiale : passage en pourcentage par ligne (n'a de sens que pour des données positives ou nulles dont la division par la somme des éléments d'une ligne est possible), suivi d'un centrage par colonne ;

2 — Pondération des lignes (3 options) : pondération uniforme (1/nombre de lignes, utilisée par défaut), pondération unitaire (1), pondération à lire dans un fichier à une colonne ;

3 — Pondération des colonnes (3 options) : pondération uniforme (1/nombre de colonnes), pondération unitaire (1, utilisée par défaut), pondération à lire dans un fichier à une colonne ;

L'option utilise une seule fenêtre de dialogue :



Nom du fichier d'entrée (binaire).

Option de pondération des lignes : 1—chaque ligne a un poids uniforme (par défaut), 2—chaque ligne a un poids unité, 3—les poids sont à lire dans un fichier à une colonne.

Option de pondération des colonnes : 1—chaque colonne a un poids uniforme, 2—chaque ligne a un poids unité (par défaut), 3—les poids sont à lire dans un fichier à une colonne.

Nom du fichier de pondération des lignes, accepté si l'option 3 de pondération des lignes est utilisée et si le fichier a autant de lignes que le fichier d'entrée et une seule colonne.

Nom du fichier de pondération des colonnes, accepté si l'option 3 de pondération des colonnes est utilisée et si le fichier a une seule colonne et autant de lignes que le fichier d'entrée a de colonnes.

On obtient les mêmes résultats en utilisant Bin->Bin : Frequencies (option pourcentage par lignes) sur le fichier d'entrée suivi de PCA : Covariance Matrix PCA. L'option ajoute une aide à l'interprétation spécifique sous forme de biplot (voir *infra*).

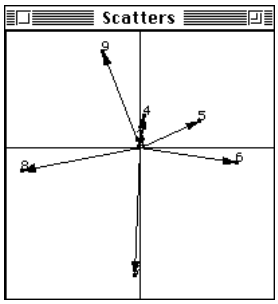
Utiliser la carte Granulo de la pile ADE-4•Data pour obtenir le fichier Granulo (49-9) :

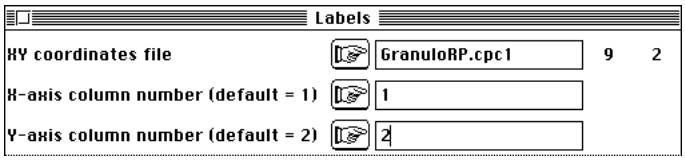
```
Row percentage PCA
Input file: Granulo
---- Row weights:
File GranuloRP.cppl contains the row weights
It has 49 rows and 1 column
Each row has 0.0204082 weight (Sum = 1)
---- Column weights:
File GranuloRP.cppc contains the column weights
It has 49 rows and 1 column
Each column has unit weight (Sum = 9)
---- Table:
File GranuloRP contains the table of row profiles (Sum by row = 1)
It has 49 rows and 9 columns
---- Table:
```

```
File GranuloRP.cpta contains the (column) centred table of row profiles
It has 49 rows and 9 columns
File :GranuloRP.cpta
|Col.|   Mini   |   Maxi   |
|----|----------|----------|
|   1|-1.491e-04| 4.518e-03|
•••
|   9|-5.755e-02| 3.829e-01|
|----|----------|----------|
---- Info: means and variances
File GranuloRP.cpma contains the descriptive of the analysis
It contains successively:
     Number of rows: 49
     Number of columns: 9
     means and variances:
     Col.: 1 | Mean: 1.4910e-04 | Variance: 4.9441e-07
•••
     Col.: 9 | Mean: 5.7551e-02 | Variance: 1.1205e-02
-----------------------------------------------------
DiagoRC: General program for two diagonal inner product analysis
Input file: GranuloRP.cpta
--- Number of rows: 49, columns: 9
-----------------------
Total inertia: 0.0652569
-----------------------
Num. Eigenval.   R.Iner.  R.Sum    |Num. Eigenval.   R.Iner.  R.Sum  |
01   +3.4211E-02 +0.5243 +0.5243   |02   +2.0315E-02 +0.3113 +0.8356 |
03   +7.1580E-03 +0.1097 +0.9453   |04   +2.8243E-03 +0.0433 +0.9885 |
05   +6.8516E-04 +0.0105 +0.9990   |06   +6.0574E-05 +0.0009 +1.0000 |
07   +2.3682E-06 +0.0000 +1.0000   |08   +0.0000E+00 +0.0000 +1.0000 |
09   +0.0000E+00 +0.0000 +1.0000

File GranuloRP.cpvp contains the eigenvalues and relative inertia for each axis
--- It has 9 rows and 2 columns

File GranuloRP.cpco contains the column scores
--- It has 9 rows and 2 columns
File :GranuloRP.cpco
|Col.|   Mini   |   Maxi   |
|----|----------|----------|
|   1|-1.294e-01| 1.070e-01|
|   2|-1.084e-01| 8.189e-02|
|----|----------|----------|

File GranuloRP.cpli contains the row scores
--- It has 49 rows and 2 columns
File :GranuloRP.cpli
|Col.|   Mini   |   Maxi   |
|----|----------|----------|
|   1|-2.995e-01| 4.250e-01|
|   2|-1.681e-01| 4.352e-01|
|----|----------|----------|

File GranuloRP.cpls contains the row scores of the non centered table
Number of rows: 49, columns: 2

File :GranuloRP.cpls
|Col.|   Mini   |   Maxi   |
|----|----------|----------|
|   1|-3.628e-01| 3.618e-01|
|   2|-4.384e-01| 1.650e-01|
|----|----------|----------|
File GranuloRP.cpc1 contains the column scores with unit norm
It has 9 rows and 2 columns
File :GranuloRP.cpc1
|Col.|   Mini   |   Maxi   |
|----|----------|----------|
|   1|-6.998e-01| 5.787e-01|
|   2|-7.603e-01| 5.745e-01|
|----|----------|----------|
```

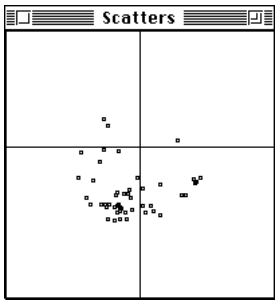
Les deux derniers fichiers sont particuliers à ce type d'analyse. Le fichier ---.cpc1 contient les poids canoniques des variables, c'est-à-dire les coefficients des variables dans les combinaisons linéaires que sont les coordonnées des lignes. Le fichier ---.cpls contient les positions des lignes par averaging sur ces scores colonnes. L'essentiel de l'information d'un plan factoriel est exprimé par un biplot (représentation simultanée des lignes et des colonnes du tableau). On croit souvent, à tort, que cette pratique est spécifique de l'analyse des correspondances. Elle dérive aisément de la théorie dont le cas particulier est la représentation triangulaire[1]. Elle a été popularisée par K.R. Gabriel[2], discutée dans [3] et utilisée dans [4]. La mise en œuvre est simple.

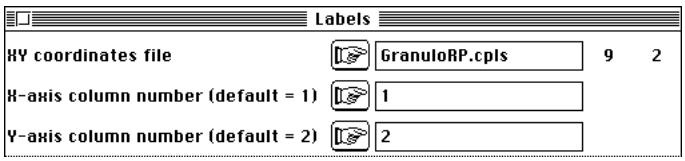
Positionner les catégories par Scatters : Labels :





Bloquer l'échelle de la fenêtre et positionner les points-lignes :





Superposer les deux dessins (Copier-Coller) dans un document MacDraw© ouvert. On obtient l'image des vecteurs de la base canonique de $\mathbb{R}^p$ (p est le nombre de variables) et celle des lignes par averaging de leur distribution. Tout se passe comme si on pouvait faire tourner la base canonique de $\mathbb{R}^p$ (ce qui pour $p>3$ n'est pas raisonnable) et qu'on s'arrêtait dans la position où le nuage des lignes présente l'image la plus lisible (variance maximale). Appliqué à trois catégories (variables), cette procédure redonne la représentation triangulaire classique.

Ce type d'analyse convient essentiellement pour des données de répartition en catégories dont le poids total est sans intérêt, par exemple les courbes granulométriques (le poids total de l'échantillon est une information peu pertinente) les contenus stomacaux (le poids de l'échantillon n'est pas maîtrisé et dépend seulement du dernier repas de l'individu capturé), les budgets temps (le poids d'une ligne dépend de la longueur de l'observation) …

Après ce module, les options de DDUtil sont disponibles.

[1] Gower, J.C. (1967) Multivariate analysis and multivariate geometry. *The statistician* : 17, 13-28.

[2] Gabriel, K.R. (1971) The biplot graphical display of matrices with application to principal component analysis. *Biometrika* : 58, 453-467.

[3] Ter Braak, C.J.F. (1983) Principal components biplots and alpha and beta diversity. *Ecology* : 64, 3, 454-462.

[4] Dolédec, S. (1986) *Les peuplements de macroinvertébrés benthiques du cours inférieur de l'Ardèche. Dynamique spatio-temporelle*. Thèse Doctorat, Univ. Lyon I. 1-246.

PCA : Correlation matrix PCA

Méthode d'analyse multivariée à un tableau.

L'Analyse en Composantes Principales (ACP) classique normée est l'analyse d'un schéma de dualité avec le jeu de paramètres :

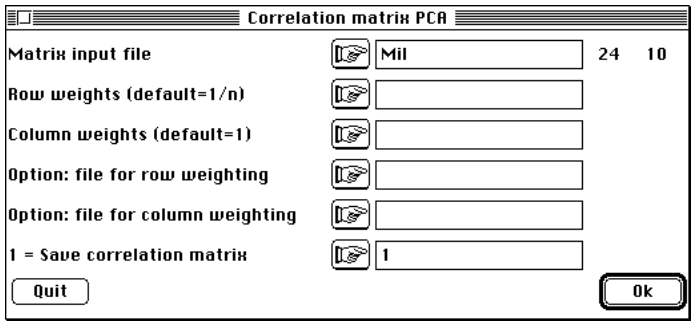
1 — Transformation initiale : normalisation par colonne. **X** est la matrice de départ, x_{ij} la valeur du terme de la ligne *i* et de la colonne *j*, *n* le nombre de lignes et *p* le nombre de colonnes. p_i est le poids associé à la ligne *i*. La normalisation est le changement de variable $x_{ij} \mapsto \left(x_{ij} - m_j\right) \Big/ s_j$, avec :

$$m_j = \sum_i p_i x_{ij} \text{ (moyenne)} \quad s_j = \sqrt{\sum_i p_i \left(x_{ij} - m_j\right)^2} \text{ (écart − type)}$$

2 — Pondération des lignes (3 options) : pondération uniforme (1/nombre de lignes, utilisée par défaut), pondération unitaire (1), pondération à lire dans un fichier à une colonne ;

3 — Pondération des colonnes (3 options) : pondération uniforme (1/nombre de colonnes), pondération unitaire (1, utilisée par défaut), pondération à lire dans un fichier à une colonne.

L'option utilise une seule fenêtre de dialogue :



Nom du fichier d'entrée (binaire).

Option de pondération des lignes : 1—chaque ligne a un poids uniforme (par défaut), 2—chaque ligne a un poids unité, 3—les poids sont à lire dans un fichier à une colonne.

Option de pondération des colonnes : 1—chaque colonne a un poids uniforme, 2—chaque ligne a un poids unité (par défaut), 3—les poids sont à lire dans un fichier à une colonne.

Nom du fichier de pondération des lignes, accepté si l'option 3 de pondération des lignes est utilisée et si le fichier a autant de lignes que le fichier d'entrée et une seule colonne.

Nom du fichier de pondération des colonnes, accepté si l'option 3 de pondération des colonnes est utilisée et si le fichier a une seule colonne et autant de lignes que le fichier d'entrée a de colonnes.

Option de sauvegarde sur fichier de la matrice de corrélation.

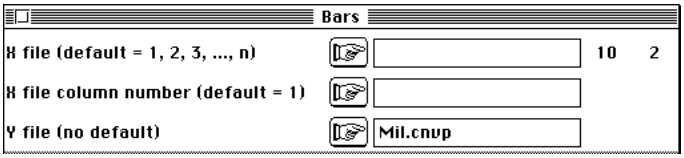
Utiliser la carte Méaudret de la pile ADE-4•Data pour obtenir le fichier Mil (24-10).

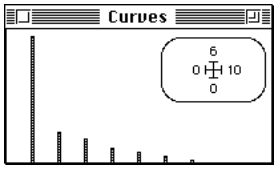
```
Classical Principal Component Analysis (Hotteling 1933)
Input file: Mil
---- Row weights:
File Mil.cnpl contains the row weights
It has 24 rows and 1 column
Each row has 0.0416667 weight (Sum = 1)
---- Column weights:
File Mil.cnpc contains the column weights
```

```
It has 24 rows and 1 column
Each column has unit weight (Sum = 10)
---- Table:
File Mil.cnta contains the centred and normed table
Zero mean and unit variance for each column
It has 24 rows and 10 columns
File :Mil.cnta
|Col.|   Mini   |   Maxi   |
|----|----------|----------|
|   1|-1.284e+00| 1.587e+00|
•••
|  10|-9.117e-01| 2.968e+00|
|----|----------|----------|

---- Info: means and variances
File Mil.cnma contains the descriptive of the analysis
It contains successively:
     Number of rows: 24
     Number of columns: 10
     means and variances:
     Col.:   1 | Mean:  7.7083e+00 | Variance: 2.7290e+01
•••
     Col.:  10 | Mean:  1.5504e+00 | Variance: 2.7809e+00
-----------------------------------------------------

File Mil.cn+r contains the Correlation matrix
from statistical triplet Mil.cnta
It has 10 rows and 10 columns

----------------------- Correlation matrix -------------------
[  1] 1000
[  2] -108 1000
[  3] -127  187 1000
[  4]  -79 -323 -716 1000
[  5] -215  354  682 -568 1000
[  6]   91 -255 -638  757 -697 1000
[  7]  110 -264 -594  765 -713  947 1000
[  8]  143 -309 -744  805 -768  963  914 1000
[  9] -116 -286   78  173  144 -198 -250 -131 1000
[ 10]   62 -330 -647  855 -646  875  814  909  229 1000
-------------------------------------------------------------

-----------------------------------------------------
DiagoRC: General program for two diagonal inner product analysis
Input file: Mil.cnta
--- Number of rows: 24, columns: 10
-----------------------
Total inertia:       10
-----------------------
Num. Eigenval.   R.Iner.  R.Sum    |Num. Eigenval.   R.Iner.  R.Sum  |
01   +5.7451E+00 +0.5745 +0.5745   |02   +1.4301E+00 +0.1430 +0.7175 |
03   +1.0838E+00 +0.1084 +0.8259   |04   +6.7612E-01 +0.0676 +0.8935 |
05   +5.2437E-01 +0.0524 +0.9459   |06   +3.0332E-01 +0.0303 +0.9763 |
07   +1.4624E-01 +0.0146 +0.9909   |08   +5.3706E-02 +0.0054 +0.9963 |
09   +2.2734E-02 +0.0023 +0.9985   |10   +1.4535E-02 +0.0015 +1.0000 |

File Mil.cnvp contains the eigenvalues and relative inertia for each axis
--- It has 10 rows and 2 columns

File Mil.cnco contains the column scores
--- It has 10 rows and 2 columns
File :Mil.cnco
|Col.|   Mini   |   Maxi   |
|----|----------|----------|
|   1|-9.781e-01| 8.141e-01|
|   2|-5.149e-01| 9.204e-01|
|----|----------|----------|

File Mil.cnli contains the row scores
--- It has 24 rows and 2 columns
File :Mil.cnli
|Col.|   Mini   |   Maxi   |
```

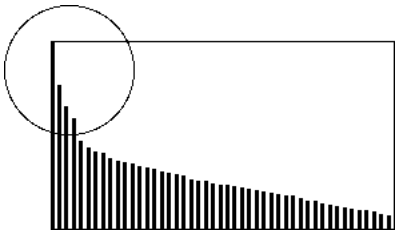
```
|----|----------|----------|
|   1|-7.234e+00| 2.332e+00|
|   2|-1.543e+00| 2.927e+00|
|----|----------|----------|
-------------------------------------------------
```

⚠ L'inertie totale des nuages de lignes ou de colonnes est égale au nombre de variables. Pour obtenir le graphe des valeurs propres, comme dans toutes les analyses à un tableau, utiliser Curves : Bars :





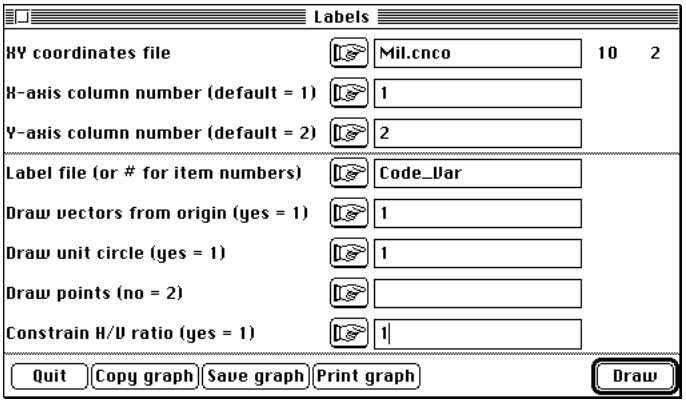
Contrairement à une croyance répandue, de forts taux d'inertie représentée sur les premiers axes ne sont pas indispensables à la pertinence d'une analyse. C'est le contraire : il est plus difficile de trouver une structure à 5% derrière un bruit de fond à 95% qu'une structure à 95% à peine perturbée. C'est la présence de premières valeurs propres nettement distinctes[1] des suivantes qui indique une vraie pertinence (ici un seul axe). Le cas le plus intéressant est du type :

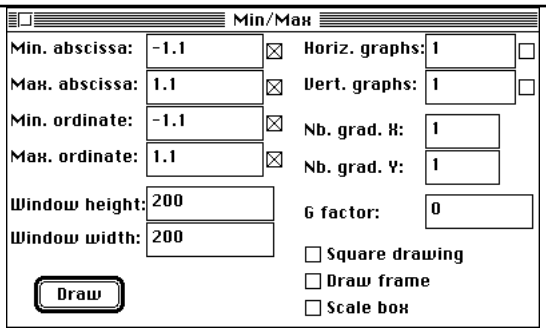


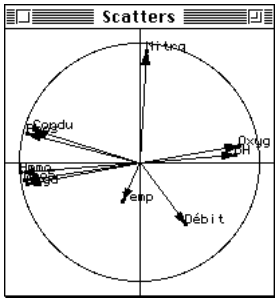
Dans ce type de circonstances fréquent, le choix du nombre de facteurs à conserver, qui continue à faire couler beaucoup d'encre, est un problème mineur. Ne pas oublier cependant qu'un facteur exprime de la corrélation. Une variable non corrélée aux autres s'exprime dans un facteur lointain de peu d'importance. Cette variable peut cependant être la plus signifiante biologiquement : la notion d'inertie la fera disparaître.

Pour savoir si on a besoin d'une telle analyse, la lecture de la fable de Ramsey[2] peut être salutaire. Utilisée pour décrire une redondance artefactuelle (liée, par exemple, à un paramètre mesurée de plusieurs manières), la méthode est de peu d'intérêt. Elle a de farouches détracteurs[3] mais elle est une des méthodes les plus utilisées pour la description synthétique de l'environnements[4] et des communautés (planctoniques[5], végétales[6], de tous les types[7]), en chimiométrie[8], morphométrie[9], géologie[10] …

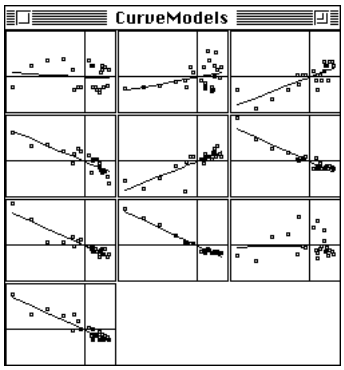
Les points variables sont sur la sphère unité et les projections se font dans un cercle dit cercle des corrélations. Utiliser Scatters : Labels :

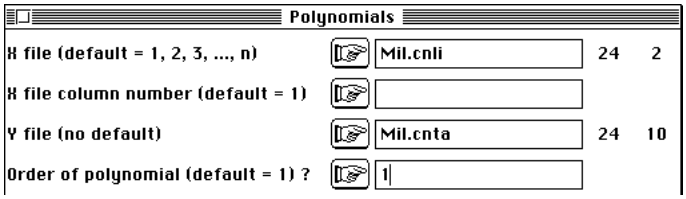






On peut voir l'analyse comme la recherche d'une variable de synthèse[11] (la première coordonnées factorielle) maximisant la somme de ses carrés de corrélation avec les variables de départ. Exprimer cette propriété avec le graphe canonique par CurveModels : Polynomial :





Depuis l'invention de Hotteling[11], le débat entre le modèle gaussien et le modèle géométrique a été sévère. On utilise ici essentiellement le second[12].

[1] Diday, E., Lemaire, J., Pouget, J. & Testu, F. (1982) Éléments d'analyse de données. Dunod, Paris. 1-462.

[2] Ramsey, F.L. (1986) A fable of PCA. *The American Statistician* : 40, 4, 323-324.

[3] Beals, E.W. (1973) Ordination : mathematical elegance and ecological naïveté. *Journal of Ecology* : 61, 23-35.

[4] Rottenberry, J.T. & Wiens, J.A. (1981) A synthetic approach to principal component analysis of bird/habitat relationships. In : *The use of multivariate statistics in studies of wildlife habitat.* USDA Forest Service General Technical Report, RM-87. Capen, D.E. (Ed.) Rocky Mountain Forest and range Experiment Station, Fort-Collins, Colorado. 197-208.

[5] Ibanez, F. (1968) Application de la méthode d'analyse en composantes principales à l'étude des populations planctoniques à l'Ouest de la Sardaigne. *Compte rendu hebdomadaire des séances de l'Académie des sciences.* Paris, D : 263, 1215-1258.

Ibanez, F. (1972) Interprétation des données écologiques par l'analyse des composantes principales : Écologie planctonique de la mer du Nord. *J. Const. int. Explor. Mer* : 34, 3, 323-340.

Frontier, S. (1974) L'analyse factorielle est-elle heuristique en écologie du plancton?. *Cahiers ORSTOM, Série Océanographie* : XII, 1, 77-81.

[6] Goodall, D.W. (1954) Objective methods for the classification of vegetation III. An essay in the use of factor analysis. *Australian Journal of Botany* : 2, 304-324.

Dagnelie, P. (1960) Contribution à l'étude des communautés végétales par l'analyse factorielle. *Bulletin du Service de la carte Phytogéographique*, B : 5, 7-71 & 93-195.

Nichols, S. (1977) On the interpretation of principal components analysis in ecological contexts. *Vegetatio* : 34, 191-197.

[7] Orloci, L. (1966) Geometric models in ecology. I. The theory and applications of some ordination methods. *Journal of Ecology* : 54, 193-215.

Austin, M.P. & Orloci, L. (1966) Geometric models in ecology II An evaluation of some ordination techniques. *Journal of Ecology* : 54, 217-227.

[8] Wold, S., Esbensen, K. & Geladi, P. (1987) Principal component analysis. *Chemometrics and Intelligent Laboratory Systems* : 2, 37-52.

Meglen, R.R. (1988) Chemometrics : its role in chemistry and measurement sciences. *Chemometrics and Intelligent Laboratory Systems* : 3, 17-29.

[9] Teissier, G. (1949) La relation d'allométrie : sa signification statistique et biologique. *Biometrics* : 4, 14-48.

Jolicœur, P. (1963) The multivariate generalization of the allometry equation. *Biometrics* : 19, 497-499.

Sprent, P. (1972) The mathematics of size and shape. *Biometrics* : 28, 23-37.

[10] Miesch, A.T. (1980) Scaling variables and interpretation of eigenvalues in principal component analysis of geological data. *Mathematical Geology* : 12, 523-538.

Webb, W.M. & Briggs, L.I. (1966) The use of principal component analysis to screen mineralogical data. *Journal of Geology* : 74, 716-720.

[11] Hotelling, H. (1933) Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* : 24, 417-441 , 498-520.

[12] Lebart, L. & Fenelon, J.P. (1971) Statistique et informatique appliquées. Dunod, Paris. 1-426.

Lebart, L., Morineau, A. & Fenelon, J.P. (1982) *Traitement des données statistiques. Méthodes et Programmes*. Dunod, 2° édition, Paris. 1-518.

Après ce module, les options de DDUtil sont disponibles.

PCA : Covariance matrix PCA

Méthode d'analyse multivariée à un tableau.

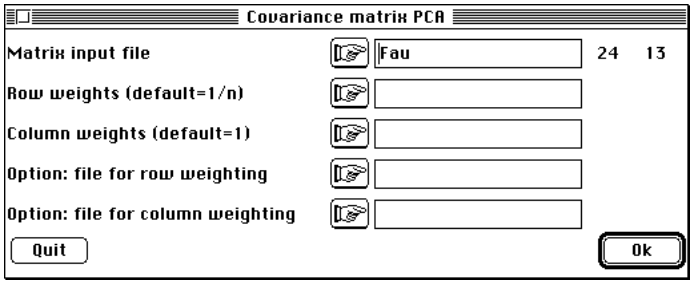
L'Analyse en Composantes Principales (ACP) classique centrée est l'analyse d'un schéma de dualité avec le jeu de paramètres :

1 — Transformation initiale : centrage par colonne. **X** est la matrice de départ, x_{ij} la valeur du terme de la ligne *i* et de la colonne *j*, *n* le nombre de lignes et *p* le nombre de colonnes. p_i est le poids associé à la ligne *i*. Le centrage est le changement de variable $x_{ij} \mapsto x_{ij} - m_j$, avec $m_j = \sum_i p_i x_{ij}$ (moyenne).

2 — Pondération des lignes (3 options) : pondération uniforme (1/nombre de lignes, utilisée par défaut), pondération unitaire (1), pondération à lire dans un fichier à une colonne ;

3 — Pondération des colonnes (3 options) : pondération uniforme (1/nombre de colonnes), pondération unitaire (1, utilisée par défaut), pondération à lire dans un fichier à une colonne.

L'option utilise une seule fenêtre de dialogue :



Nom du fichier d'entrée (binaire).

Option de pondération des lignes : 1—chaque ligne a un poids uniforme (par défaut), 2—chaque ligne a un poids unité, 3—les poids sont à lire dans un fichier à une colonne.

Option de pondération des colonnes : 1—chaque colonne a un poids uniforme, 2—chaque ligne a un poids unité (par défaut), 3—les poids sont à lire dans un fichier à une colonne.

Nom du fichier de pondération des lignes, accepté si l'option 3 de pondération des lignes est utilisée et si le fichier a autant de lignes que le fichier d'entrée et une seule colonne.

Nom du fichier de pondération des colonnes, accepté si l'option 3 de pondération des colonnes est utilisée et si le fichier a une seule colonne et autant de lignes que le fichier d'entrée a de colonnes.

Utiliser la carte Méaudret+2 de la pile ADE-4•Data pour obtenir le fichier Fau (24-13) :

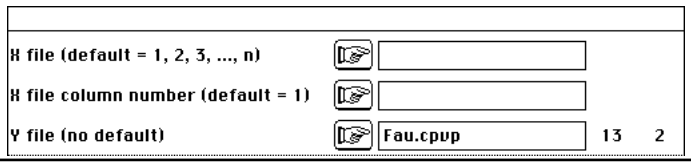
```
Centered Principal Component Analysis (Pearson 1901)
Input file: Fau
---- Row weights:
File Fau.cppl contains the row weights
It has 24 rows and 1 column
Each row has 4.1667e-02 weight (Sum = 1)
---- Column weights:
File Fau.cppc contains the column weights
It has 24 rows and 1 column
Each column has unit weight (Sum = 13)
---- Table:
File Fau.cpta contains the (column) centred table
It has 24 rows and 13 columns
File :Fau.cpta
|Col.|   Mini   |   Maxi   |
|----|----------|----------|
```

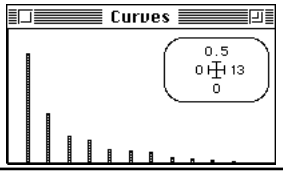
```
|   1|-9.167e-01| 5.083e+00|
•••
|  12|-4.167e-01| 4.583e+00|
|  13|-2.125e+00| 7.875e+00|
|----|----------|----------|
---- Info: means and variances
File Fau.cpma contains the descriptive of the analysis
It contains successively:
     Number of rows: 24
     Number of columns: 13
     means and variances:
     Col.:   1 | Mean: 9.1667e-01 | Variance: 2.7431e+00
     •••
     Col.:  12 | Mean: 4.1667e-01 | Variance: 1.2431e+00
     Col.:  13 | Mean: 2.1250e+00 | Variance: 9.4427e+00
-------------------------------------------------
DiagoRC: General program for two diagonal inner product analysis
Input file: Fau.cpta
--- Number of rows: 24, columns: 13
-----------------------
Total inertia:  79.7361
-----------------------
Num. Eigenval.   R.Iner.  R.Sum    |Num. Eigenval.   R.Iner.  R.Sum  |
01   +3.3048E+01 +0.4145 +0.4145   |02   +1.4861E+01 +0.1864 +0.6008 |
03   +8.3273E+00 +0.1044 +0.7053   |04   +7.0839E+00 +0.0888 +0.7941 |
05   +4.0988E+00 +0.0514 +0.8455   |06   +3.6404E+00 +0.0457 +0.8912 |
07   +3.3106E+00 +0.0415 +0.9327   |08   +1.9163E+00 +0.0240 +0.9567 |
09   +1.4861E+00 +0.0186 +0.9754   |10   +1.0154E+00 +0.0127 +0.9881 |
11   +5.5811E-01 +0.0070 +0.9951   |12   +3.4688E-01 +0.0044 +0.9995 |
13   +4.3198E-02 +0.0005 +1.0000

File Fau.cpvp contains the eigenvalues and relative inertia for each axis
--- It has 13 rows and 2 columns

File Fau.cpco contains the column scores
--- It has 13 rows and 2 columns
File :Fau.cpco
|Col.|   Mini   |   Maxi   |
|----|----------|----------|
|   1|-2.946e+00|-5.927e-02|
|   2|-1.362e+00| 2.416e+00|
|----|----------|----------|

File Fau.cpli contains the row scores
--- It has 24 rows and 2 columns
File :Fau.cpli
|Col.|   Mini   |   Maxi   |
|----|----------|----------|
|   1|-1.170e+01| 1.091e+01|
|   2|-9.184e+00| 6.827e+00|
|----|----------|----------|
```

⚠ L'analyse s'emploie pour un ensemble de variables de même unité, en particulier sur les tableaux relevés-espèces (centrage par taxon en colonne). Pour des variables d'unités différentes, la normalisation (PCA : Correlation matrix PCA) s'impose.

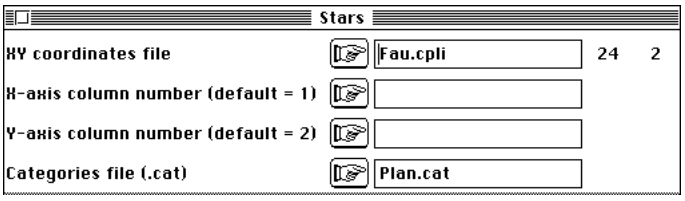
Sur les tableaux floro-faunistiques cette analyse est en concurrence avec l'analyse des correspondances. Les résultats sont souvent voisins, les deux méthodes étant contestées ensemble par les tenants des méthodes non linéaires.

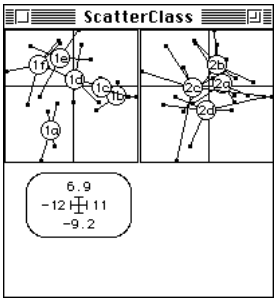
Pour dépouiller, on utilise la répartition de l'inertie — égale à somme des variances — entre les axes. Utiliser Curves : Bars sur la colonne 2 du fichier Fau.cpvp :



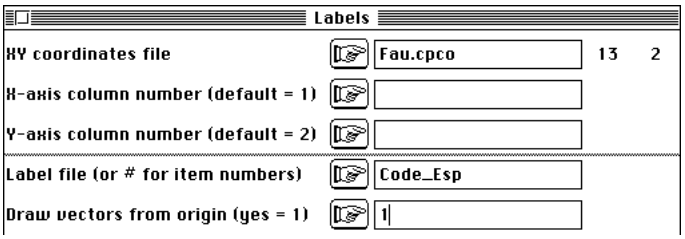


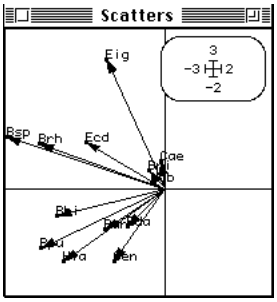
Les coordonnées des lignes sont des combinaisons linéaires de variables de variance maximale (point de vue privilégié dans [1]) en même temps que les coordonnées des projections sur les axes principaux et en ce sens sont susceptibles de toutes sortes de manipulation, comme la représentation des relevés par groupe de même station ou par groupe de même date. Utiliser le fichier Plan issu de la carte Méaudret+1 de la pile ADE-4•Data lu par CategVar : Read categ File, puis utiliser ScatterClass : Stars :





Les coordonnées des colonnes, coordonnées des projections sur le plan des composantes principales, expriment des covariances entre variables et coordonnées normalisées des lignes, covariances qui dépendent donc des variances et des corrélations. On utilise des traits pour illustrer cet aspect (Scatters : Labels) :





La position commune des taxons sur un même demi-plan traduit un “effet taille”, c’est-à-dire l’augmentation simultanée de l’abondance de tous les taxons dans certains relevés (non pollués).

L’ACP centrée dont l’invention est indiscutablement attribuée à K. Pearson[2] est considérée, aujourd’hui comme faisant partie du domaine public et s’emploie sans citation. Il en est de même de l’ACP normée (PCA : Correlation matrix PCA). Il suffit de préciser ACP sur matrice de corrélations ou ACP sur matrice de covariances pour être clair.

Le module est de logique géométrique, comme l’ensemble du logiciel ADE-4. Aucune référence pour l’utilisation de l’ACP n’est faite au modèle gaussien multivarié dans lequel les variances sont des estimations (calculée avec $1/(n-1)$) comme les axes principaux. La logique est celle de l’ACP simple (covariances) ou l’ACP standard (corrélations) du chapitre VII de [3] ou de l’ACP générale de [4]. Ceci introduit des divergences numériques légères avec le point de vue inférentiel comme celui de [5]. L’ACP est dans ADE-4 un cas particulier des méthodes d’inertie introduite en écologie par [6].

Après ce module, les options de DDUtil sont disponibles.

[1] Manly, B.F. (1994) *Multivariate Statistical Methods. A primer*. Second edition. Chapman & Hall, London. 1-215.

[2] Pearson, K. (1901) On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* : 2, 559-572.

[3] Rouanet, H. & Le Roux, B. (1993) *Analyse des données multidimensionnelles.* Dunod, paris. 1-310.

[4] Lebart, L. & Fenelon, J.P. (1971) *Statistique et informatique appliquées.* Dunod, Paris. 1-426.

Lebart, L., Morineau, A. & Tabart, N. . (1977) *Techniques de la description statistique, méthodes et logiciels pour la description des grands tableaux.* Dunod, Paris. 1-351.

[5] Legendre, L. & Legendre, P. (1984b) Tome 2 - *La structure des données écologiques.* Masson, Paris. 2ème édition revue et augmentée : 1-344.

Capen, D.E. (Ed.) (1981) *The use of multivariate statistics in studies of wildlife habitat*. USDA Forest Service General Technical Report, RM-87. Rocky Mountain Forest and range Experiment Station, Fort-Collins, Colorado. 1-249.

[6] Chardy, P., Glemarec, M. & Laurec, A. (1976) Application of inertia methods to benthic marine ecology: practical implication of the basic options. *Estuarine and Coastal Marine Science* : 4, 179-205.

Laurec, A., Chardy, P., de la Salle, P. & Rickaert, M. (1979) Use of dual structures in inertia analysis ecological implications. In : *Multivariate methods in ecological work.* Orloci, L., Rao, C.R. & Stiteler, W.M. (Eds.) Statistical Ecology Series. Vol. 7, International co-operative publishing house, Burtonsville. 127-174.

PCA : Decentring X[i,j] - Model[i,j]

Méthode d'analyse multivariée à un tableau.

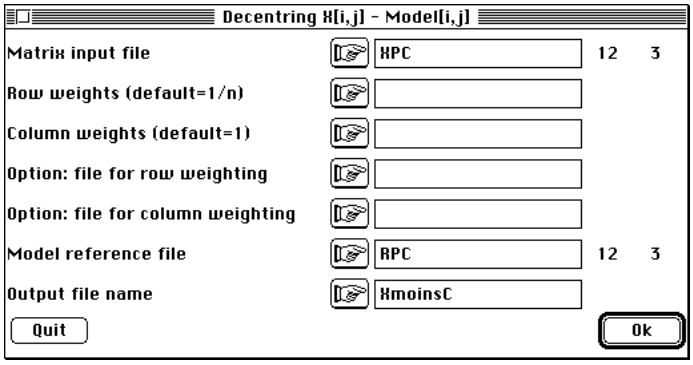
L'Analyse en Composantes Principales (ACP) décentrée sur un tableau-modèle est l'analyse d'un schéma de dualité avec le jeu de paramètres :

1 — Transformation initiale : décentrage sur un tableau. **X** est la matrice de départ, x_{ij} la valeur du terme de la ligne *i* et de la colonne *j*, *n* le nombre de lignes et *p* le nombre de colonnes. **R** est la matrice de référence, r_{ij} la valeur du terme de la ligne *i* et de la colonne *j*, *n* le nombre de lignes et *p* le nombre de colonnes. Le décentrage est la transformation $x_{ij} \mapsto x_{ij} - r_{ij}$.

2 — Pondération des lignes (3 options) : pondération uniforme (1/nombre de lignes, utilisée par défaut), pondération unitaire (1), pondération à lire dans un fichier unicolonne ;

3 — Pondération des colonnes (3 options) : pondération uniforme (1/nombre de colonnes), pondération unitaire (1, utilisée par défaut), pondération à lire dans un fichier unicolonne.

L'option utilise une seule fenêtre de dialogue :



Nom du fichier d'entrée (binaire).

Option de pondération des lignes : 1—chaque ligne a un poids uniforme (par défaut), 2—chaque ligne a un poids unité, 3—les poids sont à lire dans un fichier unicolonne.

Option de pondération des colonnes : 1—chaque colonne a un poids uniforme, 2—chaque ligne a un poids unité (par défaut), 3—les poids sont à lire dans un fichier unicolonne.

Nom du fichier de pondération des lignes, accepté si l'option 3 de pondération des lignes est utilisée et si le fichier a autant de lignes que le fichier d'entrée et une seule colonne.

Nom du fichier de pondération des colonnes, accepté si l'option 3 de pondération des colonnes est utilisée et si le fichier a une seule colonne et autant de lignes que le fichier d'entrée a de colonnes.

Nom du fichier du tableau de référence.

Nom générique des fichiers de sortie.

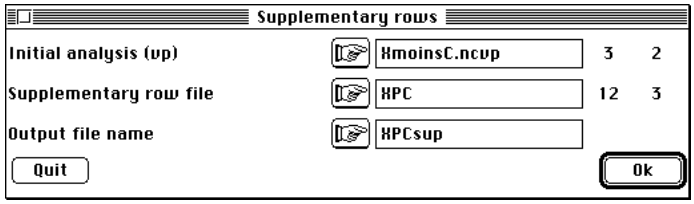
Utiliser la carte Europe de la pile ADE-4•Data pour obtenir le fichier Euro12/6 (12-6). Extraire les colonnes 1, 2 et 3 dans R (12-3) et les colonnes 4, 5 et 6 dans X (12-3) par FilesUtil : Row-Col Selection. Passer les deux fichiers en pourcentage par lignes (Bin->Bin : Frequencies) dans XPC (12-3) et RPC (12-3).

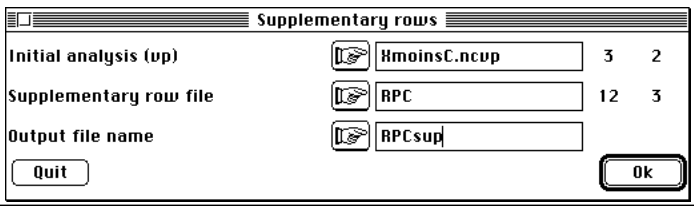
Utiliser la présente option :

```
Principal Component Analysis using a model
Input file: XPC
```

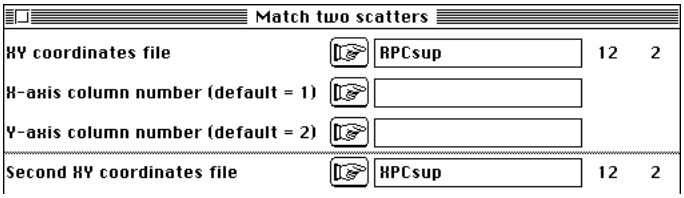
```
---- Row weights:
File XmoinsC.ncpl contains the row weights
It has 12 rows and 1 column
Each row has 8.3333e-02 weight (Sum = 1)
---- Column weights:
File XmoinsC.ncpc contains the column weights
It has 12 rows and 1 column
Each column has unit weight (Sum = 3)
---- Table:
File XmoinsC.ncta contains the table Data-Model
with reference model from file RPC
It has 12 rows and 3 columns
File :XmoinsC.ncta
|Col.|   Mini   |   Maxi   |
|----|----------|----------|
|   1|-9.600e-02|-3.000e-03|
|   2|-8.100e-02| 0.000e+00|
|   3| 4.300e-02| 9.700e-02|
|----|----------|----------|
-------------------------------------------------
DiagoRC: General program for two diagonal inner product analysis
Input file: XmoinsC.ncta
--- Number of rows: 12, columns: 3
-----------------------
Total inertia: 0.0103833
-----------------------
Num. Eigenval.   R.Iner.  R.Sum    |Num. Eigenval.   R.Iner.  R.Sum  |
01   +9.3141E-03 +0.8970 +0.8970   |02   +1.0692E-03 +0.1030 +1.0000 |
03   +0.0000E+00 +0.0000 +1.0000

File XmoinsC.ncvp contains the eigenvalues and relative inertia for each
axis
--- It has 3 rows and 2 columns

File XmoinsC.ncco contains the column scores
--- It has 3 rows and 2 columns
File :XmoinsC.ncco
|Col.|   Mini   |   Maxi   |
|----|----------|----------|
|   1|-4.780e-02| 7.815e-02|
|   2|-2.123e-02| 2.464e-02|
|----|----------|----------|

File XmoinsC.ncli contains the row scores
--- It has 12 rows and 2 columns
File :XmoinsC.ncli
|Col.|   Mini   |   Maxi   |
|----|----------|----------|
|   1| 5.467e-02| 1.185e-01|
|   2|-8.235e-02| 4.156e-02|
|----|----------|----------|
```

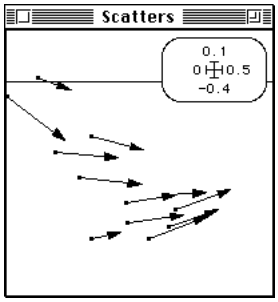
⚠ Utiliser DDUtil : Supplementary rows pour projeter en individus supplémentaires les lignes des tableaux de départ :



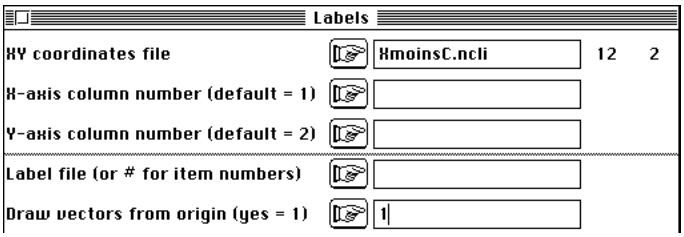


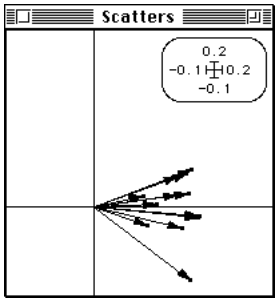
L'opération permet de caractériser l'objectif particulier de cette analyse. Utiliser Scatters : Match two scatters sur les coordonnées supplémentaires :





Puis utiliser Scatters : Labels sur le fichier des coordonnées des lignes :





Dans l'espace R^p la ligne de rang *i* du tableau **R** (référence) définit un point de départ D_i et dans ce même espace la ligne de rang *i* du tableau **X** (observations) définit un point d'arrivée A_i. Les vecteurs liés D_iA_i voient leur origine ramenée à l'origine par le décentrage et l'analyse cherche l'axe principal de ce nouveau nuage. Cet axe est celui de l'évolution entre **R** et **X** (si **R** est le tableau avant et **X** est le tableau après). L'axe d'évolution (la convergence imposée par la tertiarisation aux pays d'Europe) est distinct de l'axe typologique (l'étalement des pays européens en terme d'effondrement de l'agriculture).

Cet exercice montre deux faits très importants :

1 — L'ACP est un principe de base, très général, de représentation des données multivariées par projection sur des plans optimaux. Ce principe est unificateur et dérive d'un modèle mathématique unique (schéma de dualité[1]).

2 — On peut appliquer ce principe général d'un très grand nombre de manières particulières. Certains de ces cas particuliers sont bien connus : ACP centrée ou ACP normée. D'autres ne le sont pas du tout.

Concrètement, le paramètre le plus sensible est la transformation initiale, qui modifie sensiblement le point de vue et fait fortement varier les résultats obtenus. Les 8 options du module sont ainsi 8 manières d'utiliser le même principe, qui donne son nom au module, de façons variées, qui donnent leur nom aux options.

C'est la confusion introduite par l'usage du même terme (ACP, Analyse en Composantes Principales) pour le principe général et pour chaque cas particulier de mise en œuvre du principe qui complique la tâche des utilisateurs. Les logiciels classiques confondent le principe et deux de ses applications (mêmes unités : ACP centrée ou unités différentes : ACP normée). Pour des variantes, la citation du principe et la définition de la transformation initiale sont indispensables.

Après ce module, les options de DDUtil sont disponibles.

1 Escoufier, Y. (1987) The duality diagramm : a means of better practical applications. In : *Development in numerical ecology*. Legendre, P. & Legendre, L. (Eds.) NATO advanced Institute , Serie G .Springer Verlag, Berlin. 139-156.

PCA : Decentring X[i,j] - Model[j]

Méthode d'analyse multivariée à un tableau.

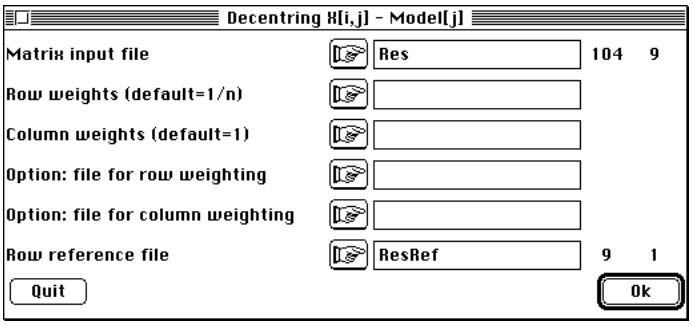
L'Analyse en Composantes Principales (ACP) décentrée sur un point de référence est l'analyse d'un schéma de dualité avec le jeu de paramètres :

1 — Transformation initiale : décentrage sur un vecteur ligne. **X** est la matrice de départ, x_{ij} la valeur du terme de la ligne *i* et de la colonne *j*, *n* le nombre de lignes et *p* le nombre de colonnes. **r** est le vecteur de référence, r_j la valeur du terme de rang *j* ($1 \le j \le p$). Le décentrage est la transformation $x_{ij} \mapsto x_{ij} - r_j$.

2 — Pondération des lignes (3 options) : pondération uniforme (1/nombre de lignes, utilisée par défaut), pondération unitaire (1), pondération à lire dans un fichier unicolonne ;

3 — Pondération des colonnes (3 options) : pondération uniforme (1/nombre de colonnes), pondération unitaire (1, utilisée par défaut), pondération à lire dans un fichier unicolonne.

L'option utilise une seule fenêtre de dialogue :



Nom du fichier d'entrée (binaire).

Option de pondération des lignes : 1—chaque ligne a un poids uniforme (par défaut), 2—chaque ligne a un poids unité, 3—les poids sont à lire dans un fichier unicolonne.

Option de pondération des colonnes : 1—chaque colonne a un poids uniforme, 2—chaque ligne a un poids unité (par défaut), 3—les poids sont à lire dans un fichier unicolonne.

Nom du fichier de pondération des lignes, accepté si l'option 3 de pondération des lignes est utilisée et si le fichier a autant de lignes que le fichier d'entrée et une seule colonne.

Nom du fichier de pondération des colonnes, accepté si l'option 3 de pondération des colonnes est utilisée et si le fichier a une seule colonne et autant de lignes que le fichier d'entrée a de colonnes.

Nom du fichier du vecteur de référence.

Utiliser la carte Deug de la pile ADE-4•Data pour obtenir les fichiers Res (104-9) et ResRef (9-1).

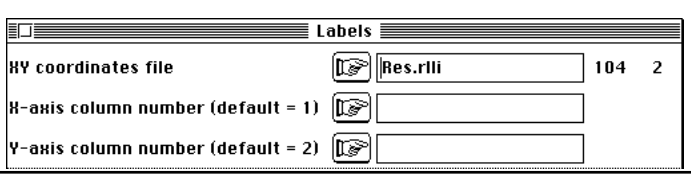
```
Row decentered Principal Component Analysis
Input file: Res
---- Row weights:
File Res.rlpl contains the row weights
It has 104 rows and 1 column
Each row has 9.6154e-03 weight (Sum = 1)
---- Column weights:
File Res.rlpc contains the column weights
It has 104 rows and 1 column
Each column has unit weight (Sum = 9)
---- Table:
```

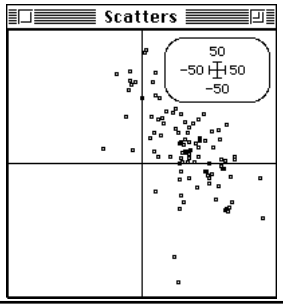
```
File Res.rlta contains the decentered table
with reference point from file ResRef
It has 104 rows and 9 columns
File :Res.rlta
|Col.|   Mini   |   Maxi   |
|----|----------|----------|
|   1|-4.100e+01| 2.200e+01|
|   2|-1.400e+01| 2.800e+01|
•••
|   9| 0.000e+00| 1.500e+01|
|----|----------|----------|
-----------------------------------------------------
DiagoRC: General program for two diagonal inner product analysis
Input file: Res.rlta
--- Number of rows: 104, columns: 9
-----------------------
Total inertia:  927.216
-----------------------
Num. Eigenval.   R.Iner.  R.Sum    |Num. Eigenval.   R.Iner.  R.Sum  |
01   +4.0696E+02 +0.4389 +0.4389   |02   +2.6377E+02 +0.2845 +0.7234 |
03   +7.0346E+01 +0.0759 +0.7992   |04   +5.8958E+01 +0.0636 +0.8628 |
05   +4.6838E+01 +0.0505 +0.9133   |06   +3.4162E+01 +0.0368 +0.9502 |
07   +2.0617E+01 +0.0222 +0.9724   |08   +1.5302E+01 +0.0165 +0.9889 |
09   +1.0263E+01 +0.0111 +1.0000

File Res.rlvp contains the eigenvalues and relative inertia for each axis
--- It has 9 rows and 2 columns

File Res.rlco contains the column scores
--- It has 9 rows and 2 columns
File :Res.rlco
|Col.|   Mini   |   Maxi   |
|----|----------|----------|
|   1|-1.388e+01| 8.740e+00|
|   2| 1.460e+00| 9.219e+00|
|----|----------|----------|

File Res.rlli contains the row scores
--- It has 104 rows and 2 columns
File :Res.rlli
|Col.|   Mini   |   Maxi   |
|----|----------|----------|
|   1|-1.432e+01| 4.536e+01|
|   2|-4.470e+01| 4.217e+01|
|----|----------|----------|
```

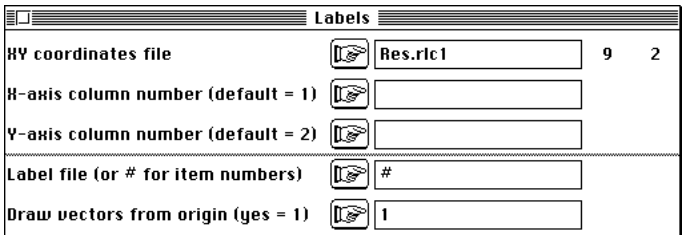
⚠ Dans une ACP centrée classique, le nuage des points-lignes a son centre de gravité à l'origine. On se place au centre du nuage pour voir quelles sont ses directions principales. On décide, ici, de se placer en un point particulier. Dans l'exemple, le point de référence est celui qui passionnent les étudiants (10/20 dans chaque matière). Le tableau décentré est celui qui compte matière par matière les points "perdus" ou "gagnés" pour réussir l'examen.

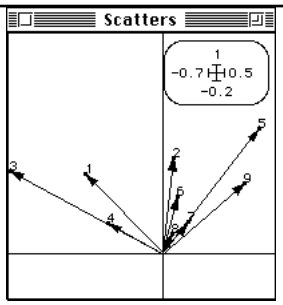
La carte factorielle des lignes place ce point de référence à l'origine et décrit le nuage par rapport à ce point. Utiliser DDUtil : Add normed scores pour obtenir la projection des vecteurs de la base canonique de $\mathbb{R}^p$ sur les axes principaux (coordonnées de normes unité).

Employer Scatters : Labels pour représenter les résultats.









On obtient un biplot décentré, image du nuage des 109 étudiants et des 9 matières. A droite, observer les “profs sympas” qui poussent le nuage du coté des points positifs (5 est celui d’éducation physique qui ne donne que des points de bonification). A gauche, noter la présence des “profs de maths” qui étalent le nuage.

Le centrage habituel est un point de vue et cette option permet d’en avoir un autre. On retiendra l’idée fondamentale sous le nom d’analyse générale ([1] page 283) : “un tableau peut donner lieu à des représentations sous forme de nuages de points dans deux espaces, et les ajustements de ces deux nuages sont liés par des relations simples”.

On trouvera un autre exemple de décentrage signifiant sur la carte Sport&Ville de la pile ADE-4•Data.

Une version AFC du décentrage est disponible par COA : Decentred COA.

Après ce module, les options de DDUtil sont disponibles.

[1] Lebart, L., Morineau, A. & Fenelon, J.P. (1982) *Traitement des données statistiques. Méthodes et Programmes.* Dunod, 2° édition, Paris. 1-518.

PCA : Non centred PCA

Méthode d'analyse multivariée à un tableau.

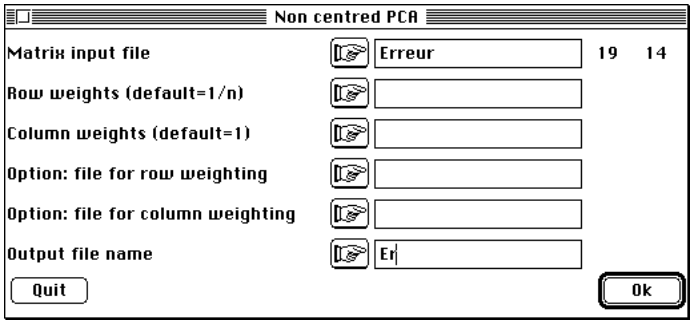
L'Analyse en Composantes Principales (ACP) non centrée est l'analyse d'un schéma de dualité avec le jeu de paramètres :

1 — Transformation initiale : aucune.

2 — Pondération des lignes (3 options) : pondération uniforme (1/nombre de lignes, utilisée par défaut), pondération unitaire (1), pondération à lire dans un fichier unicolonne ;

3 — Pondération des colonnes (3 options) : pondération uniforme (1/nombre de colonnes), pondération unitaire (1, utilisée par défaut), pondération à lire dans un fichier unicolonne.

L'option utilise une seule fenêtre de dialogue :



Nom du fichier d'entrée (binaire).

Option de pondération des lignes : 1—chaque ligne a un poids uniforme (par défaut), 2—chaque ligne a un poids unité, 3—les poids sont à lire dans un fichier unicolonne.

Option de pondération des colonnes : 1—chaque colonne a un poids uniforme, 2—chaque ligne a un poids unité (par défaut), 3—les poids sont à lire dans un fichier unicolonne.

Nom du fichier de pondération des lignes, accepté si l'option 3 de pondération des lignes est utilisée et si le fichier a autant de lignes que le fichier d'entrée et une seule colonne.

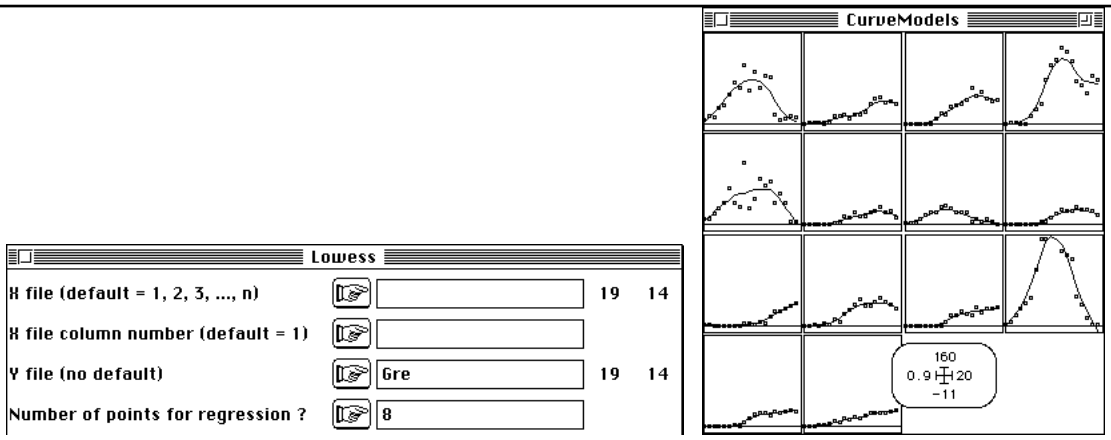
Nom du fichier de pondération des colonnes, accepté si l'option 3 de pondération des colonnes est utilisée et si le fichier a une seule colonne et autant de lignes que le fichier d'entrée a de colonnes.

Nom générique des fichiers de sortie.

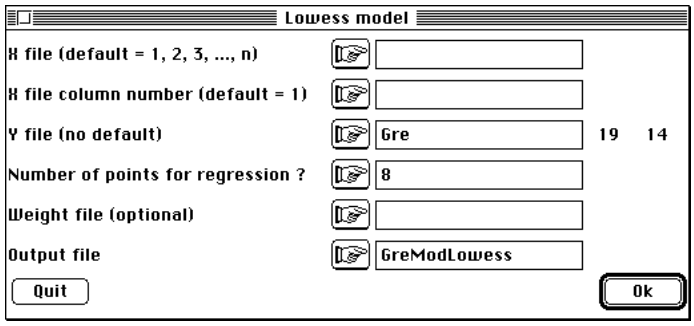
Utiliser la carte Grèbes de la pile ADE-4•Data pour obtenir les fichiers Gre (19-14). Il contient 14 courbes enregistrées en 19 sites. Une ACP centrée par sites positionnerait chaque courbe par rapport à la courbe constante. Une ACP centrée par date positionnerait chaque site par rapport à un site moyen.

Ne sortiraient de ces opérations que des évidences puisque les rythmes d'évolution en chaque site s'expriment clairement et que les quantités globales (nombre de couples nicheurs) sont manifestement très différents entre sites (étangs).

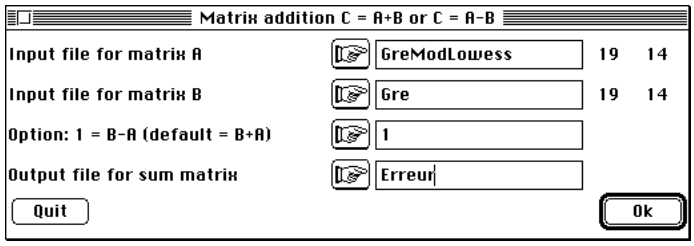
Utiliser CurveModels : Lowess pour voir les données :



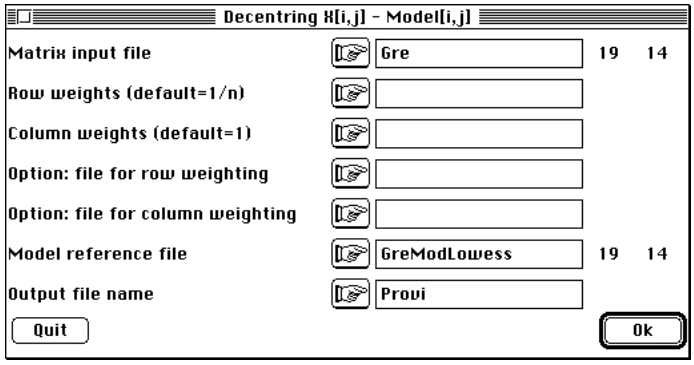
Récupérer ces modèles satisfaisants dans GraphUtils : Lowess model :



Calculer les résidus dans MatAlg : Matrix addition C = A+B or C = A-B :



Vérifier que l'usage de la présente option est équivalente à :



⚠ L'absence de transformation initiale permet à cette option d'exécuter une quelconque des versions de l'analyse de données à un tableau. Toute modification du tableau initial peut être assurée par les fonctions de base d'un tableur.

Après ce module, les options de DDUtil sont disponibles.

La généralité du modèle euclidien d'analyse des données a été complètement maîtrisée et décrite, en écologie, par I. Noy-Meir (1973)[1] qui décrit 12 méthodes d'ordination linéaire (dont l'AFC) d'un tableau floro-faunistique. La clarté et la précision des articles de l'auteur justifient leur citation lors de tout usage d'une version de l'ACP autre que les standards centrée ou normée dont il a été le premier à contester l'unicité.

[1] Noy-Meir, I. & Austin M.P. (1970) Principal component ordination and simulated vegetational data. *Ecology*: 51, 551-552.

Noy-Meir, I. (1971) Multivariate analysis of the semi-arid vegetation in South-eastern Australia. I. Nodal ordination by component analysis. *Proceedings of Ecological society of Australia*:: 6, 159-193.

Noy-Meir, I. & Anderson D.J. (1971) Multivariate pattern analysis, or multiscale ordination: towards a vegetation hologram ? In : Patil, G.P., Pielou, E.C. & Waters, W.E. (1971) *Statistical Ecology, III Many species populations ecosystems and systems analysis*. Pennsylvania State University Press: 208-231.

Noy-Meir, I. (1973) Data transformations in ecological ordination. I. Some advantages of non-centering. *Journal of Ecology* : 61,329-341.

Noy-Meir, I. (1974) Multivariate analysis of the semi-arid vegetation in South-eastern Australia. II. Vegetation Catenae and environmental gradients. *Australian Journal of Botany* : 22, 115-140.

Noy-Meir, I. (1974) Catenation: quantitative methods for the definition of coenoclines. *Vegetatio*: 29, 89-99.

Noy-Meir, I., Walker, D. & Williams, W.T. (1975) Data transformation in ecological ordination. II On the meaning of data standardization. *Journal of Ecology* : 63, 779-800.

Noy-Meir, I. & Whittaker, R.H. (1977) Continuous multivariate methods in community analysis: some problems and developments. *Vegetatio* : 33, 79-98.

Noy-Meir, I. & van der Maarel, E. (1987) Relations between community theory and community analysis in vegetation science: some historical perspectives. *Vegetatio* : 69, 5-15.

PCA : Normed Y[i,j] - X[i,j]

Méthode d'analyse multivariée à un tableau.

L'Analyse en Composantes Principales (ACP) des différences normée est l'analyse d'un schéma de dualité avec le jeu de paramètres :

1 — Transformation initiale : décentrage sur un tableau et normalisation. **X** est une matrice de données, x_{ij} la valeur du terme de la ligne i et de la colonne j, n le nombre de lignes et p le nombre de colonnes. **Y** est une autre matrice de données portant sur les mêmes lignes-individus et les mêmes colonnes-variables, y_{ij} la valeur du terme de la ligne i et de la colonne j. On calcule la norme des différences pour chacune des variables :

$$\frac{1}{n}\sum_{i=1,n}\left(y_{ij}-x_{ij}\right)^2=\partial(j)$$

Le tableau analysé est **Z** de terme général :

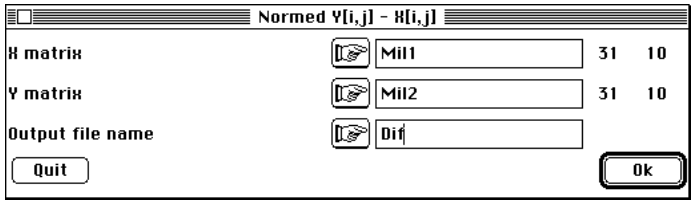
$$z_{ij}=\frac{y_{ij}-x_{ij}}{\sqrt{\partial(j)}}$$

2 — La pondération des lignes est uniforme (1/nombre de lignes) ;

3 — La pondération des colonnes est unitaire (1).

Chaque vecteur des différences (en colonne dans **Z**) est un vecteur normé de R^n et justifie une ACP normée implicite.

L'option utilise une seule fenêtre de dialogue :



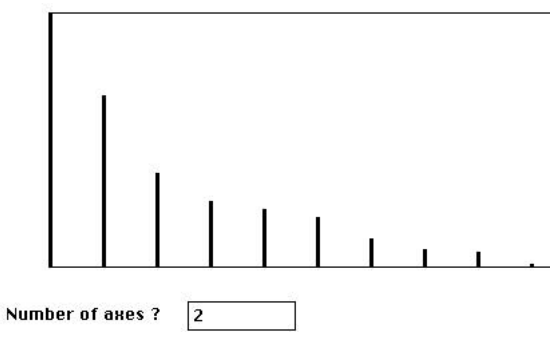
Nom du fichier du tableau **X**.

Nom du fichier du tableau **Y**.

Nom générique des fichiers de sortie.

Utiliser la carte Buech 1 2 de la pile ADE-4•Data. Le premier tableau Mil1 contient 31 mesures sur 10 variables physiques et chimiques. Il est acquis en juin. Le second tableau Mil2 donne les valeurs des mêmes variables sur les mêmes individus. Il est acquis en septembre.

Analyser la différence des deux tableaux par la présente option :



Le matériel utilisé est rappelé :

```
X Table: Mil1
It has 31 rows and 10 columns
```

```
Y Table: Mil2
It has 31 rows and 10 columns
Generic output file name: Dif
---- Row weight:
File Dif.cnpl contains the row weight
It has 31 rows and 1 column
Each row has 3.2258e-02 weight (Sum = 1)
---- Column weights:
File Dif.cnpc contains the column weights
It has 10 rows and 1 column
Each column has unit weight (Sum = 10)
---- Table:
File Dif.cnta contains the table Y-X
It has 31 rows and 10 columns
```

Le triplet statistique est diagonalisé. L'inertie totale, comme dans une ACP normée est égale au nombre de variables :

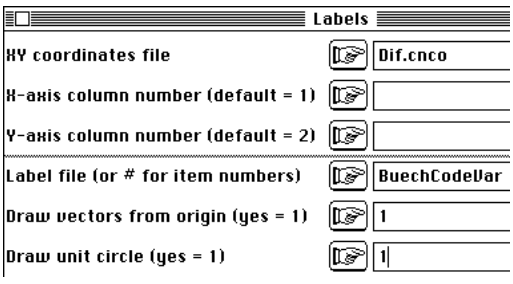
```
DiagoRC: General program for two diagonal inner product analysis
Input file: Dif.cnta
--- Number of rows: 31, columns: 10
-----------------------
Total inertia:       10
-----------------------
Num. Eigenval.   R.Iner.  R.Sum    |Num. Eigenval.   R.Iner.  R.Sum  |
01   +3.3400E+00 +0.3340 +0.3340   |02   +2.2502E+00 +0.2250 +0.5590 |
03   +1.2381E+00 +0.1238 +0.6828   |04   +8.7669E-01 +0.0877 +0.7705 |
05   +7.6389E-01 +0.0764 +0.8469   |06   +6.6381E-01 +0.0664 +0.9133 |
07   +3.8168E-01 +0.0382 +0.9514   |08   +2.3289E-01 +0.0233 +0.9747 |
09   +2.1035E-01 +0.0210 +0.9958   |10   +4.2334E-02 +0.0042 +1.0000 |
```

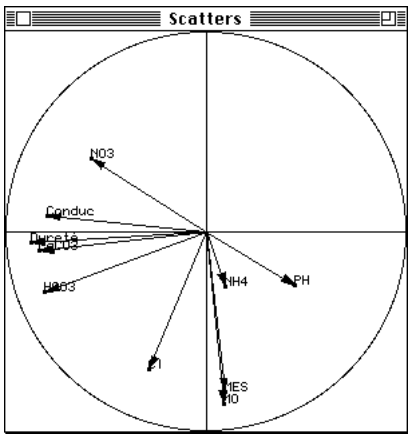
Les valeurs propres sont conservées :

```
File Dif.cnvp contains the eigenvalues and relative inertia for each axis
--- It has 10 rows and 2 columns
```

Les vecteurs des différences normalisés sont projetés sur les composantes principales. On obtient les points à l'intérieur d'un cercle :

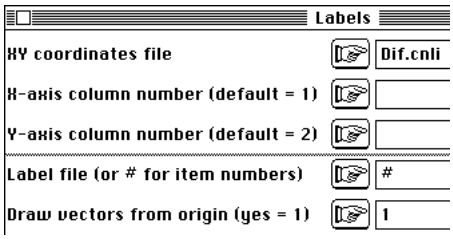
```
File Dif.cnco contains the column scores
--- It has 10 rows and 2 columns
File :Dif.cnco
|Col.|   Mini   |   Maxi   |
|----|----------|----------|
|   1|-8.707e-01| 4.419e-01|
|   2|-8.622e-01| 3.606e-01|
|----|----------|----------|
```

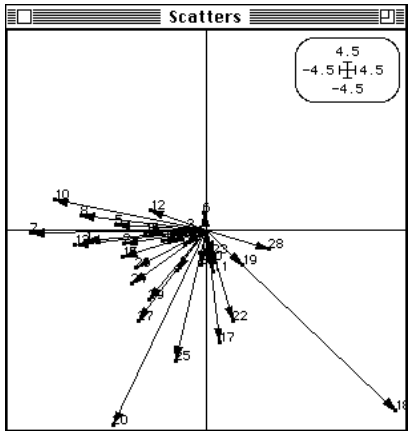





```
File Dif.cnli contains the row scores
--- It has 31 rows and 2 columns
File :Dif.cnli
|Col.|   Mini   |   Maxi   |
|----|----------|----------|
|   1|-3.938e+00| 4.284e+00|
|   2|-4.395e+00| 6.609e-01|
|----|----------|----------|
```

Les lignes du tableau des différences normalisées sont projetées sur les axes principaux :





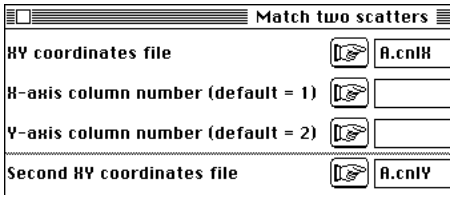
L'origine est la représentation d'un point dans le tableau **X** et l'extrémité est la représentation du même point dans le tableau **Y**. Reste à l'origine un point ayant strictement la même position dans les deux tableaux. L'extrémité caractérise l'évolution d'une station entre les deux saisons.

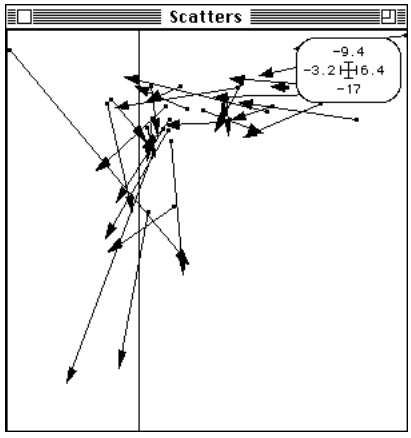
La projection en individus supplémentaires des données transformées ayant généré par différence les axes factoriels, soit

$$\frac{x_{ij}^1}{\sqrt{\partial(j)}} \quad \text{et} \quad \frac{x_{ij}^2}{\sqrt{\partial(j)}}$$

renforce cette appréciation :

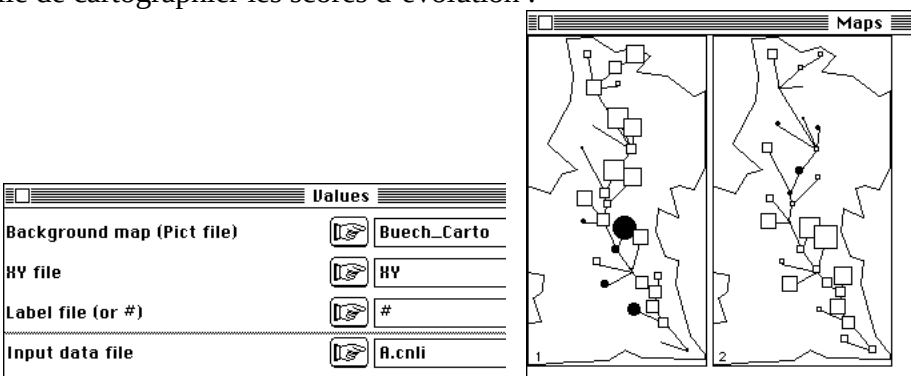
```
File A.cnlX contains the row scores of the non centered table
Number of rows: 31, columns: 2
File :A.cnlX
|Col.|   Mini   |   Maxi   |
|----|----------|----------|
|   1|-3.114e+00| 6.385e+00|
|   2|-1.287e+01|-9.523e+00|
|----|----------|----------|

File A.cnlY contains the row scores of the non centered table
Number of rows: 31, columns: 2
File :A.cnlY
|Col.|   Mini   |   Maxi   |
|----|----------|----------|
|   1|-1.741e+00| 3.683e+00|
|   2|-1.611e+01|-9.779e+00|
|----|----------|----------|
```



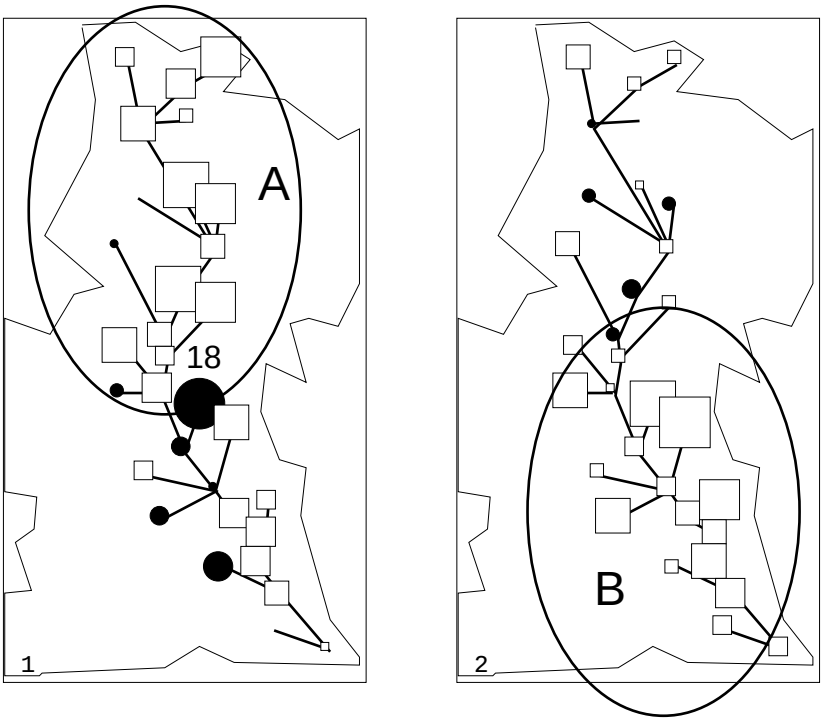


C'est l'opération qui consiste à replacer l'origine des vecteurs à l'origine des axes qui génère l'analyse d'inertie, comme dans une analyse des différences ordinaires (PCA : Decentring X[i,j] - Model[i,j]) mais les unités hétérogènes entre variables impose ici l'usage de PCA : Normed Y[i,j] - X[i,j].

On pense alors que les types d'évolution des variables eux-mêmes sont régionalisés et il est utile de cartographier les scores d'évolution :



Voir l'interprétation dans la fiche thématique 4-6.



Après ce module, les options de DDUtil sont disponibles.

Cette analyse s'applique au couple de tableaux amont-aval [3] ou avant-après pour des données d'unités hétérogènes.

[1] Vespini, F. (1985) *Contribution à l'étude hydrobiologique du Buech, rivière non aménagée de Haute-Provence*. Thèse de troisième cycle, Université de Provence. 1-148 + bibliographie + annexes.

[2] Vespini, F., Légier, P. & Champeau, A. (1987) Ecologie d'une rivière non aménagée des Alpes du Sud : Le Buëch (France) I Evolution longitudinale des descripteurs physiques et chimiques. *Annales de Limnologie* : 23, 151-164.

[3] Castella, E., Bickerton, M., Armitage, P.D. & Petts, G.E. (1995) The effects of water abstractions on invertebrate communities in U.K. streams. *Hydrobiol ogia* : 308, 167-182.

PCA : Partial normed PCA

Méthode d'analyse multivariée à un tableau.

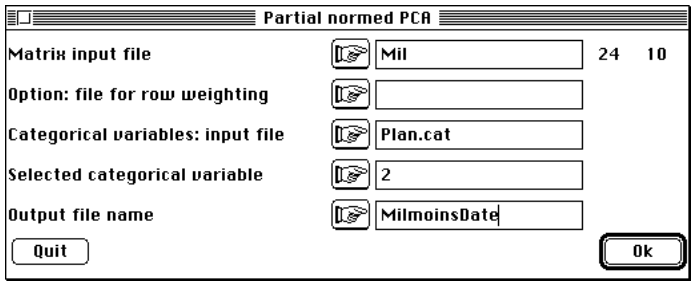
L'Analyse en Composantes Principales (ACP) partielle normée est l'analyse d'un schéma de dualité avec le jeu de paramètres :

1 — Transformation initiale : centrage par bloc de lignes suivie d'une normalisation par colonne.

2 — Pondération des lignes (2 options) : pondération uniforme (1/nombre de lignes, utilisée par défaut), pondération à lire dans un fichier unicolonne ;

3 — Pondération des colonnes unitaire (1).

L'option utilise une seule fenêtre de dialogue :



Nom du fichier d'entrée (binaire).

Option de pondération des lignes : si un nom de fichier unicolonne est utilisé, les poids sont à lire dans ce fichier. Par défaut ils sont uniformes.

Nom de fichier de type ---.cat qui permet l'accès à un fichier de variables qualitatives (après CategVar : Read Categ file).

Numéro de la variable sélectionnée dans le fichier de variables qualitatives pour définir les groupes de lignes.

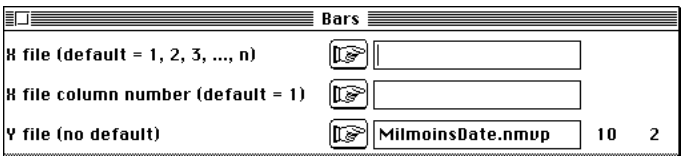
Nom générique des fichiers de sortie.

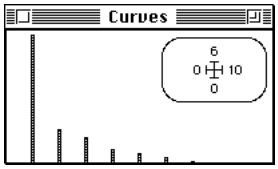
Utiliser la carte Méaudret de la pile ADE-4•Data pour obtenir le fichier Mil (24-10), puis récupérer le fichier Plan issu de la carte Méaudret+1 et le lire par CategVar : Read categ File.

```
Within class PCA analysis
Bouroche, J.M. (1975) Analyse des données ternaires:
La double analyse en composantes principales.
Thèse de 3° cycle, Université de Paris VI. 1-57 + annexes.

Input file: Mil
Categories defined by column 2 of file identified by Plan.cat

File MilmoinsDate.nmpc contains the column weights (equal to 1)
It has 10 rows and 1 column

File MilmoinsDate.nmpl contains the row weights
It has 24 rows and 1 column

File MilmoinsDate.nmta contains the block-centered and normalized by
columns
It has 24 rows and 10 columns

-----------------------------------------------------
DiagoRC: General program for two diagonal inner product analysis
Input file: MilmoinsDate.nmta
--- Number of rows: 24, columns: 10
```

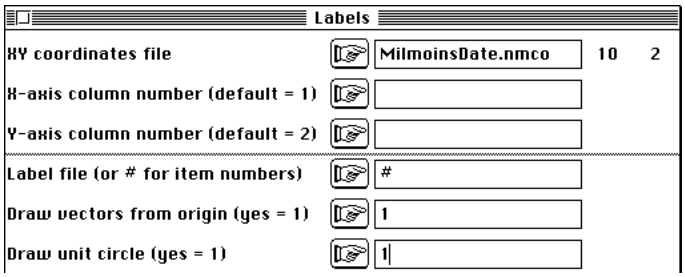
```
-----------------------
Total inertia:       10
-----------------------
Num. Eigenval.   R.Iner.  R.Sum    |Num. Eigenval.   R.Iner.  R.Sum  |
01   +5.8057E+00 +0.5806 +0.5806   |02   +1.5455E+00 +0.1545 +0.7351 |
03   +1.1757E+00 +0.1176 +0.8527   |04   +6.0494E-01 +0.0605 +0.9132 |
05   +4.4992E-01 +0.0450 +0.9582   |06   +2.8040E-01 +0.0280 +0.9862 |
07   +6.5373E-02 +0.0065 +0.9928   |08   +4.2340E-02 +0.0042 +0.9970 |
09   +1.8501E-02 +0.0019 +0.9988   |10   +1.1658E-02 +0.0012 +1.0000 |

File MilmoinsDate.nmvp contains the eigenvalues and relative inertia for
each axis
--- It has 10 rows and 2 columns

File MilmoinsDate.nmco contains the column scores
--- It has 10 rows and 2 columns
File :MilmoinsDate.nmco
|Col.|   Mini   |   Maxi   |
|----|----------|----------|
|   1|-9.804e-01| 8.071e-01|
|   2|-2.111e-02| 8.504e-01|
|----|----------|----------|

File MilmoinsDate.nmli contains the row scores
--- It has 24 rows and 2 columns
File :MilmoinsDate.nmli
|Col.|   Mini   |   Maxi   |
|----|----------|----------|
|   1|-7.764e+00| 2.784e+00|
|   2|-2.090e+00| 2.198e+00|
|----|----------|----------|
```

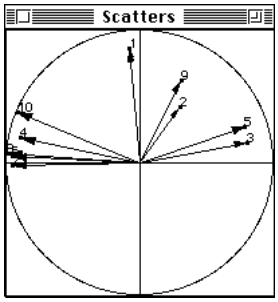
⚠ L'inertie totale est égale au nombre de variables, car la logique de l'analyse est celle d'une ACP normée (vecteurs colonnes sur la sphère unité). Les valeurs propres sont représentées par Curves : Bars sur la colonne 1 du fichier Fau.cpvp :



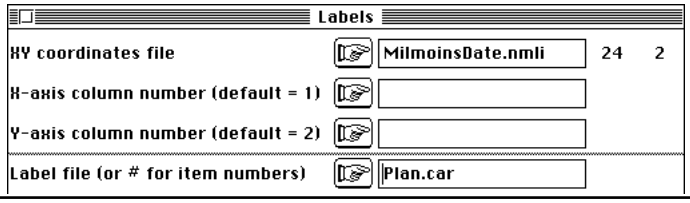


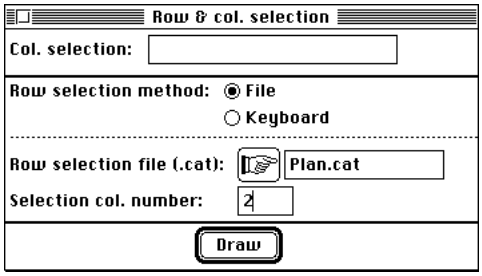
La carte des colonnes est un cercle de corrélations, comme dans une ACP normée, mais il s'agit de corrélations partielles débarrassée de l'effet de la variable qualitative qui définit les groupes (Scatters : Labels) :

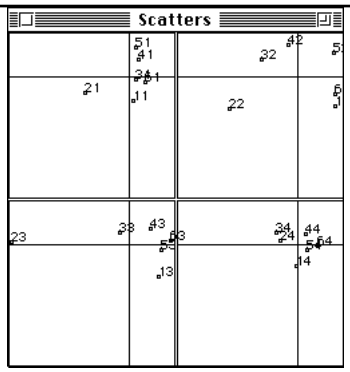




La carte des lignes est centrée par classe, car on a affaire à une analyse intra-classe :







Ce type d'analyse s'utilise comme une ACP normé, pour des données quantitatives avec unités hétérogènes entre variables, mais substitue la corrélation partielle (sachant l'effet groupe) à la corrélation habituelle. Pour la notion de corrélation partielle, consulter [1] (p. 300 et suivantes). L'option correspond à l'analyse des corrélations partielles de cet ouvrage. Elle fait partie des propositions de [2] sur les données spatio-temporelles. Elle diffère d'une intra-classe ordinaire (module Discrimin) car elle n'élimine pas les variables à forte variance inter-classe et elle diffère d'une intra-classe normée par classe (PCA : Within group normalized PCA) en maintenant les différences de variabilité dans chaque classe.

Après ce module, les options de DDUtil sont disponibles.

[1] Lebart, L., Morineau, A. & Fenelon, J.P. (1982) Traitement des données statistiques. Méthodes et Programmes. Dunod, 2° édition, Paris. 1-518.

[2] Bouroche, J.M. (1975) Analyse des données ternaires: la double analyse en composantes principales. Thèse de 3° cycle, Université de Paris VI. 1-57 + annexes.

PCA : Within group normalized PCA

Méthode d'analyse multivariée à un tableau.

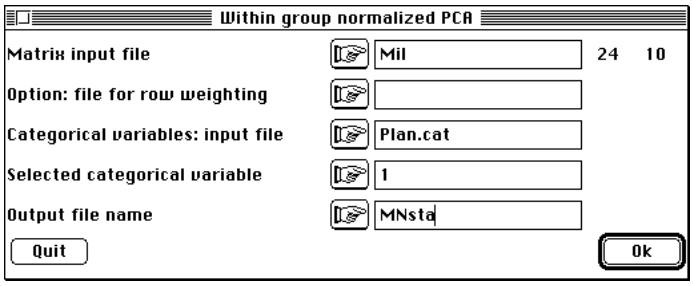
L'Analyse en Composantes Principales (ACP) partielle normée par bloc est l'analyse d'un schéma de dualité avec le jeu de paramètres :

1 — Transformation initiale : normalisation par bloc de lignes.

2 — Pondération des lignes (2 options) : pondération uniforme (1/nombre de lignes, utilisée par défaut), pondération à lire dans un fichier unicolonne ;

3 — Pondération des colonnes unitaire (1).

L'option utilise une seule fenêtre de dialogue :



Nom du fichier d'entrée (binaire).

Option de pondération des lignes : si un nom de fichier unicolonne est utilisé, les poids sont à lire dans ce fichier. Par défaut ils sont uniformes.

Nom de fichier de type ---.cat qui permet l'accès à un fichier de variables qualitatives (après CategVar : Read Categ file).

Numéro de la variable sélectionnée dans le fichier de variables qualitatives pour définir les groupes de lignes.

Nom générique des fichiers de sortie.

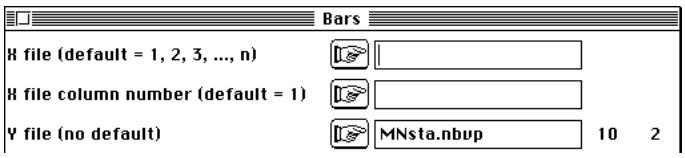
Utiliser la carte Méaudret de la pile ADE-4•Data pour obtenir le fichier Mil (24-10), puis récupérer le fichier Plan issu de la carte Méaudret+1 et le lire par CategVar : Read categ File.

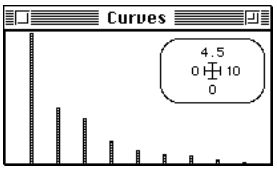
```
Within class (block normalization) PCA analysis
Input file: Mil
Categories defined by column 1 of file identified by Plan.cat

File MNsta.nbpc contains the column weights (equal to 1)
It has 10 rows and 1 column

File MNsta.nbpl contains the row weights
It has 24 rows and 1 column

File MNsta.nbta contains the block-centered and normalized by columns
It has 24 rows and 10 columns

-----------------------------------------------------
DiagoRC: General program for two diagonal inner product analysis
Input file: MNsta.nbta
--- Number of rows: 24, columns: 10
-----------------------
Total inertia:       10
-----------------------
Num. Eigenval.   R.Iner.  R.Sum    |Num. Eigenval.   R.Iner.  R.Sum  |
01   +4.4170E+00 +0.4417 +0.4417   |02   +1.9349E+00 +0.1935 +0.6352 |
03   +1.5444E+00 +0.1544 +0.7896   |04   +7.8371E-01 +0.0784 +0.8680 |
05   +4.5969E-01 +0.0460 +0.9140   |06   +3.1740E-01 +0.0317 +0.9457 |
```

```
07   +2.7902E-01 +0.0279 +0.9736   |08   +1.6246E-01 +0.0162 +0.9899 |
09   +6.5236E-02 +0.0065 +0.9964   |10   +3.6200E-02 +0.0036 +1.0000 |

File MNsta.nbvp contains the eigenvalues and relative inertia for each
axis
--- It has 10 rows and 2 columns

File MNsta.nbco contains the column scores
--- It has 10 rows and 3 columns
File :MNsta.nbco
|Col.|   Mini   |   Maxi   |
|----|----------|----------|
|   1|-7.492e-01| 8.544e-01|
|   2|-8.942e-01| 7.883e-01|
|   3|-5.937e-01| 7.189e-01|
|----|----------|----------|

File MNsta.nbli contains the row scores
--- It has 24 rows and 3 columns
File :MNsta.nbli
|Col.|   Mini   |   Maxi   |
|----|----------|----------|
|   1|-3.293e+00| 4.366e+00|
|   2|-2.482e+00| 2.262e+00|
|   3|-2.277e+00| 2.428e+00|
|----|----------|----------|
```

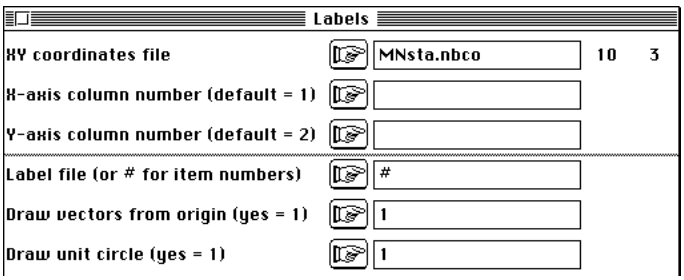
⚠ L'inertie totale est égale au nombre de variables, car la normalisation par blocs implique la normalisation totale. Une exception est faite si une variable est constante dans un bloc. Dans ce cas, toutes les valeurs correspondantes sont mises à 0 et la variance globale de la variable est diminuée de l'inverse du nombre de blocs. On pourra vérifier que l'analyse intra-date correspondante présent une inertie totale de 9.75 (4 blocs).

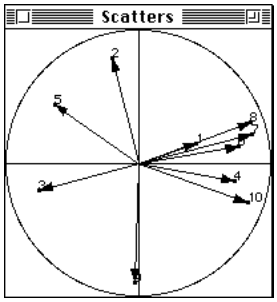
Les valeurs propres sont représentées par Curves : Bars sur la colonne 1 du fichier Fau.cpvp :



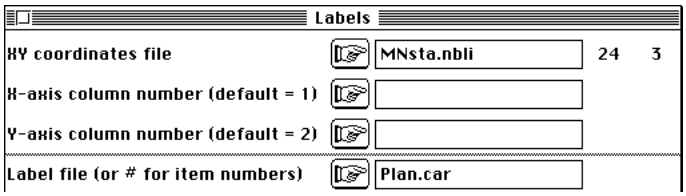


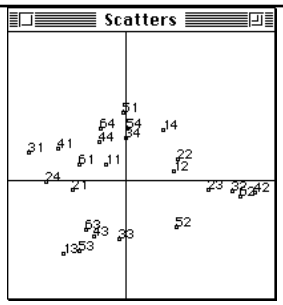
La carte des colonnes est un cercle de corrélations, comme dans une ACP normée, mais c'est une conséquence de la préparation des données. Une variance de 1 par bloc fait une variance totale de 1 puisque la variance inter-classe est ramenée à 0 par le recentrage. La représentation d'un cercle de corrélation est inféodé à la norme unité des variables mais les cosinus ne sont plus des corrélations classiques mais des moyennes de corrélations intra-classes (Scatters : Labels) :

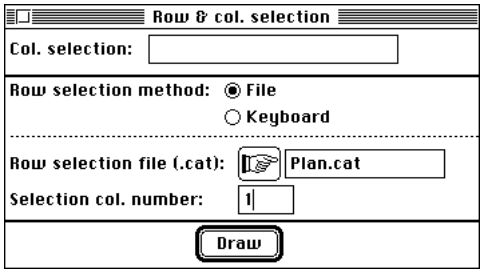


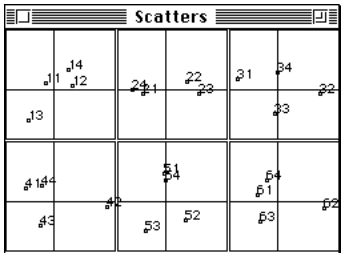


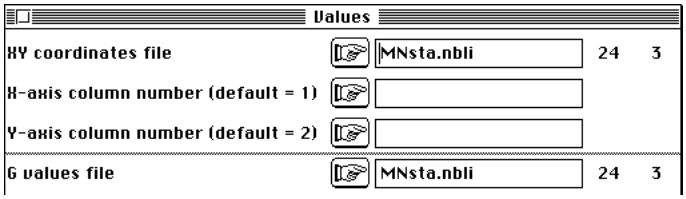
La carte des lignes (Scatters) peut être globale, répartie en classes (cette analyse intra-classe a des coordonnées des lignes centrées par classe), ou représenter le troisième facteur dans le plan des deux premiers (voir [1]) :

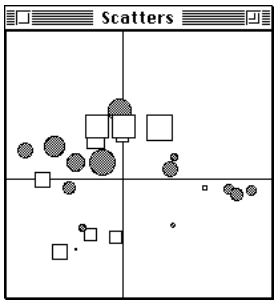












Ce type d'analyse s'utilise pour des données quantitatives dont on veut stabiliser la variance par classe et homogénéiser la participation des variables. Elle diffère d'une intra-classe ordinaire (module Discrimin) car elle n'élimine pas les variables à forte variance inter-classe et d'une ACP partielle normée (PCA : Partiel normed PCA) car elle unifie la variabilité dans chaque classe. Elle est utilisée dans [1].

Après ce module, les options de DDUtil sont disponibles.

[1] Dolédec, S. & Chessel, D. (1987) Rythmes saisonniers et composantes stationnelles en milieu aquatique I - Description d'un plan d'observations complet par projection de variables. *Acta Œcologica, Œcologia Generalis* : 8, 3, 403-426..